SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Refined Feature Selection Scheme-Based Machine Learning Model With SMOTE And Hyperparameter Tuning Approach For Diabetes Prediction

Rahul Ranjan[1] , Mohd Haroon[2]

[1]Department of CSE Integral University, Lucknow, Indiarrtiwari88@gmail.com
[2]Department of CSE Integral University, Lucknow, India mharoon@iul.ac.in

ABSTRACT
Diabetes mellitus is a syndrome that runs with lifelong metabolism that is affecting numerous individuals and requires high precision in early diagnosis in order to minimize complications and related expenses. This paper is suggested a novel hybrid framework for classification of diabetes that applies Recursive Feature Elimination (RFE), Synthetic Minority Oversampling Technique (SMOTE), hyperparameter tuning and SVM learning. In the initial stages the phase of data preprocessing applied over the Pima Indian Diabetes Dataset (PIDD) is by executing method of mean imputation and min-max normalization. It helped to increase the data information quality. Then, the RFE algorithm is applied for extraction of the significant features from the dataset: Blood Pressure, Glucose, Body Mass Index (BMI) and Insulin. For solving the purpose of the class imbalance issue, the SMOTE technique is applied prior to modeling, whereas the hyperparameter tuning is performed using Grid Search method. Proposed model is compared with several machine learning classifiers, such as Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbor (KNN), Histogram Gradient Boosting (HGB), and Naïve Bayes (NB). Experimental results have shown that the suggested RFE-SMOTE-SVM model with hyperparameter tuning demonstrates the best performance and reaches 93.1% accuracy, 92.5% precision, 92.4% recall, 91.1% F1-score and AUC equal to 0.97. Comparative analysis proves that the suggested framework outperforms existing approaches by resolving feature redundancy, class imbalance and classifier optimization issues.

KEYWORDS: Diabetes detection, Machine learning, Recursive feature elimination (RFE), SMOTE, Hyperparameter tuning.

INTRODUCTION

The main duty of public health organisations is to safeguard communities against illnesses and other health dangers. In order to achieve this goal, governments devote a large portion of their GDP to public health efforts, such as vaccination campaigns, which have greatly increased average life expectancy [1]. Due to its detrimental effects on the neurological, renal, and cardiovascular systems, diabetes mellitus is a serious health problem [2]. Furthermore, the illness affects almost every organ in the human body [3] and is frequently made worse by conditions including smoking, obesity, and high blood pressure. Millions of people are impacted by this global health issue, which emphasises the significance of early diagnosis and effective treatment methods [4]. Stroke, cardiovascular conditions, macrovascular illnesses, and other long-term vascular consequences are linked to chronic diabetes [5].

Determining a person's risk of acquiring chronic conditions like diabetes is crucial since early detection can improve treatment outcomes and save medical costs. Even when patients are unconscious or unable to respond, doctors must frequently make appropriate judgements in life-threatening medical circumstances [6]. Many statistical, analytical, and predictive methods are used for diagnosis and result predictions in order to support such clinical judgements. Computers may learn patterns from data without explicitly programmed instructions thanks to machine learning (ML), a well-known subfield of artificial intelligence (AI) [7]. Machine learning (ML) has greatly enhanced computer systems' prediction powers through quick and effective processing of massive amounts of data. It is used in many different fields, including as illness detection, pattern

recognition, and picture categorisation. Models are trained through supervised learning techniques, and their outputs are subsequently assessed using various performance metrics to evaluate their effectiveness [8]. Machine learning serves both as a feature selection method and a classification tool. It improves diagnostic accuracy while addressing a major challenge in diabetes research, namely the efficient classification of individuals for risk assessment. Numerous ML-based algorithms have been employed for diabetes prediction and classification applications [9]. The vast amount of information generated within modern healthcare systems provides significant opportunities for predictive analysis. When properly utilized, these datasets can reveal valuable information regarding disease trends and associated risk factors. Data obtained from the long-term Pima Indian study conducted in Arizona has been extensively used in diabetes research and forms the basis for many predictive models. Furthermore, advanced AI-driven approaches have been developed to apply data mining techniques for the early detection of diabetes [10].

The proposed framework integrates machine learning with recursive feature elimination (RFE) to develop an efficient diabetes classification model that leverages the advantages of feature selection and sequential data learning while reducing challenges such as gradient vanishing in deep learning systems. RFE serves as an initial filtering mechanism by repeatedly ranking feature importance and eliminating less significant attributes. Through this process, the dataset is refined to four key predictors—glucose, insulin, blood pressure, and BMI—thereby reducing complexity and improving model interpretability. The reduced feature set simplifies the learning procedure and enhances computational efficiency. Following the removal of redundant variables, the gradient decay mitigation mechanism enables computational resources to focus more effectively on identifying temporal relationships and hidden patterns within the data.

A machine learning algorithm based on a selected feature set can effectively identify patterns within diabetes-related datasets. Its architecture is well suited for modelling such relationships because it controls the flow of information through update and reset mechanisms, enabling the selective retention or removal of information according to temporal changes. A common limitation of conventional machine learning approaches is the vanishing gradient problem, which occurs when gradients become progressively smaller during backpropagation, making it difficult to learn long-term dependencies. This challenge is addressed through a gated mechanism that maintains stable gradient flow throughout the training process. In particular, the update gate determines how much information from previous states should be preserved while regulating the incorporation of new inputs. By minimizing gradient decay, the architecture improves the classifier's ability to learn and capture long-term dependencies within the data.

The integration of RFE with gradient decay mitigation techniques reduces redundant data processing that commonly results from high-dimensional datasets and irrelevant features. Through the selection of the most informative attributes, RFE decreases noise and complexity, thereby improving the stability of gradient propagation within the ML framework. By addressing gradient stability challenges while extracting critical features for diabetes classification, the proposed hybrid approach develops a robust and efficient model that enhances prediction accuracy and accelerates convergence. Furthermore, the RFE-ML model addresses important research limitations by effectively handling class imbalance, improving feature selection, and achieving better computational efficiency than many existing approaches. In contrast, several earlier models were unable to remove redundant features effectively, which frequently resulted in overfitting and reduced generalization capability.

The RFE technique enhances model performance by systematically selecting the most relevant features, reducing dimensionality, and improving overall efficiency. Traditional models often encounter challenges associated with class imbalance in the PIDD dataset; however, the RFE-ML framework addresses this issue by concentrating on discriminative features, thereby improving precision and recall, particularly for minority classes. In addition, the gradient decay control mechanism prevents both vanishing and exploding gradients, leading to improved classification performance when processing sequential data. Many existing models also fail to capture temporal dependencies effectively. Although k-fold cross-validation was not employed, its application could provide additional validation of the model's robustness and further improve the generalizability of the framework across different healthcare datasets. Overall, the RFE-ML model provides a more accurate, efficient, and reliable approach for diabetes classification, addressing limitations of earlier studies and enhancing predictive performance on small and imbalanced datasets such as PIDD.

The following summarizes this study's main contributions:

For diabetes classification, this study employs different machine learning techniques by integrating RFE with gradient decay mitigation. A sequential hybrid RFE-ML model is developed to identify and retain the most

relevant features from the dataset used for model training, thereby improving prediction performance for the target variable. In addition, the proposed model addresses common deep learning challenges, such as vanishing and exploding gradients, particularly in the features selected through RFE. The remainder of the paper is organised as follows. Section 2 presents a review of relevant studies related to diabetes prediction. Section 3 provides a detailed description of the research methodology, including data preprocessing and the applied machine learning models. Section 4 describes the proposed RFE-ML model, whereas Section 5 presents the results obtained from various experiments. Finally, Section 6 discusses and presents the findings of the study.

RELATED WORK

Diabetes diagnosis and classification is a complex task that requires specialized medical knowledge. Recent developments in healthcare, supported by ML, ensemble learning, and deep learning techniques, have significantly enhanced disease prediction capabilities. Several methods have been proposed for diabetes classification, with many of them utilizing PIDD, a dataset containing health records of female patients based on nine features [9]. This section presents a review of important studies related to diabetes classification approaches. Butt et al. [2] proposed an ML-based approach for diabetes detection, classification, and prediction. Three classifiers, namely Random Forest (RF), Logistic Regression (LR), and Multilayer Perceptron (MLP), were evaluated. Data analysis-based prediction was also conducted on PIDD using Linear Regression (LR), Moving Average (MA), and Long Short-Term Memory (LSTM) methods. Among the evaluated classifiers, MLP achieved the highest prediction accuracy of 86.08%, which was further increased to 87.26% using LSTM.

Kumari and Kaur [4] applied several supervised ML learning techniques, including ANN, linear SVM, RBF kernel SVM, multifactor dimensionality reduction (MDR), and k-nearest neighbor (k-NN), to the PIDD dataset. Among the evaluated methods, linear SVM demonstrated the best performance for diabetes prediction, achieving a maximum accuracy of 89%. Krishnamoorthi et al. [8] proposed an ML-based framework for diabetes prediction by evaluating RF and SVM models. The proposed approach produced a relatively low error rate and achieved an accuracy of 83%. Hrimov et al. [9] worked for introducing a LR-based diagnostic model for estimation of the risk of diabetes. The result shows an accuracy of 77.06% based on Python-based programming. In a research work the study carried out by Yahyaoui et al. [11] ML and DL techniques applied for designing a diabetes prediction system. CNN and SVM methods were applied for detection of diseases. The predictions with RF outperformed SVM and DL techniques with maximum accuracy of 83.67%.

In a study, Sharma et al. [12] focused on the Hrimov et al. [9] presented a LR-based diagnostic model for estimating diabetes risk, which achieved an accuracy of 77.06% through a Python-based implementation. Yahyaoui et al. [11] integrated ML and DL approaches to develop a diabetes prediction system. CNN and SVM were employed for disease detection and prediction, while RF achieved better performance than both SVM and DL models, attaining the highest accuracy of 83.67%.

Sharma et al. [12] focused on diabetes prognosis by applying supervised learning methods, including NB, LR, DT, and ANN. Among the evaluated approaches, LR achieved the highest performance with an accuracy of 80.43%. Khanam and Foo [13] employed ML, data mining, and NN techniques for diabetes prediction. Among the seven ML models examined, the combination of SVM and LR provided the best results. Several NN architectures were also investigated, and the network consisting of two hidden layers achieved the highest accuracy of 89%. Hassan et al. [14] emphasized the importance of early diabetes diagnosis by using supervised learning techniques such as NB, LR, DT, and ANN. Among these mentioned techniques, LR found to give highest performance at an accuracy of 80.43%. Khanam and Foo [13] implemented ML, data mining, and NN techniques for the prediction of diabetes. They investigated for seven ML techniques and the results proved that the combination of SVM and LR performed better than any other technique. Additionally, various NN structures also tested, and the one with two hidden layers gave highest accuracy of 89%. Hassan et al. [14] discussed the significance of early detection of diabetes.

They collected data from the KDC containing 13 variables and 289 instances. The proposed model achieved an accuracy of 88%, whereas both XGB and RF obtained an accuracy of 86.36%. Howlader et al. [15] applied ML models for T2D classification using different FS and classification techniques. Among the evaluated methods, the GBR model produced the highest performance, achieving an accuracy of 90.91%.

Çalisir and Dogantekin [16] introduced an automatic diabetes diagnosis method with LDA-MWSVM classifier. They also used the PIDD dataset. Kaardaan and Dadgar [17] proposed a hybrid approach that combines the a 2-layer NN with UTA algorithm for prediction of diabetes. The method, which employed FS via UTA and neural network plus genetic evolution scheme-based prediction on the PIDD dataset. Chen et al. [18] proposed a ML

model for T2D prediction with an accuracy of 90.04% using PIDD dataset. They used a J48 classifier and K-means for dimensionality reduction. Haritha et al. [19] suggested a hybrid ML scheme that combined the Cuckoo Search (CS) and Firefly algorithms. The achieved accuracy was 81% using KNN and Cuckoo fuzzy KNN on the PIMA dataset.

Zhang et al. [20] proposed an MLFFNN using the PIDD dataset to develop diabetes risk indicators. After training the network with the LM method, an accuracy of more than 80% was achieved. Benbelkacem and Atmani [21] employed an RF classifier containing 100 trees for diabetes classification using the PIDD dataset and obtained accuracies ranging from 70% to 80%. Khanwalkar and Soni [22] performed diabetes prediction on the PIDD dataset using an SMO approach based on quadratic programming. Patra and Kuntia [23] proposed an SDKNN method using the PIDD dataset, where distance calculation was performed using the standard deviation values of KNN features. The improved weighted SDKNN method achieved an accuracy of 83.76%.

Bhoi et al. [24] utilized several supervised learning algorithms, including CT, RF, NN, KNN, SVM, AB, NB, and LR, for diabetes prediction using the Pima Indian women's clinical PIDD dataset. Ramesh et al. [25] proposed an end-to-end healthcare monitoring system for diabetes management and risk assessment based on the PIDD dataset. The proposed model obtained accuracy, sensitivity, and specificity values of 83.20%, 87.20%, and 79.00%, respectively. Salem et al. [26] applied a data preprocessing strategy involving FS, outlier reduction, normalization, and missing-value imputation to the PIDD dataset. By optimizing the hyperparameters through grid search, their fuzzy-KNN classifier incorporating membership uncertainty achieved an accuracy of 90.63%.

METHODOLOGY

This section is presenting a brief description about the technical advancements that are related with the development of diabetes prediction tools. The presented classifier o is designed mainly for T2D patients, incorporating the patient input based on the PIDD dataset. First, PIDD is gathered, after which exploratory data analysis (EDA) is conducted in order to analyze the properties of available data sources. Next, data preprocessing is performed to eliminate outliers, duplicates, and missing values. The cleaned dataset is used for selecting and training of appropriate machine learning models. At last, developed models are tested and compared in terms of accuracy, sensitivity, and specificity. A general view of the proposed approach is shown in Figure 1.

Data preprocessing

Missing value imputation and outlier detection are frequently used data preprocessing methods that help to improve the quality of raw datasets. Many researches have already been done about the significance of different data preprocessing techniques, including missing value imputation, outlier detection, data reduction, data transformation, data normalization, and data segmentation [27]. In the present research, missing value imputation and data normalization procedures are used for preprocessing of the PIDD dataset.

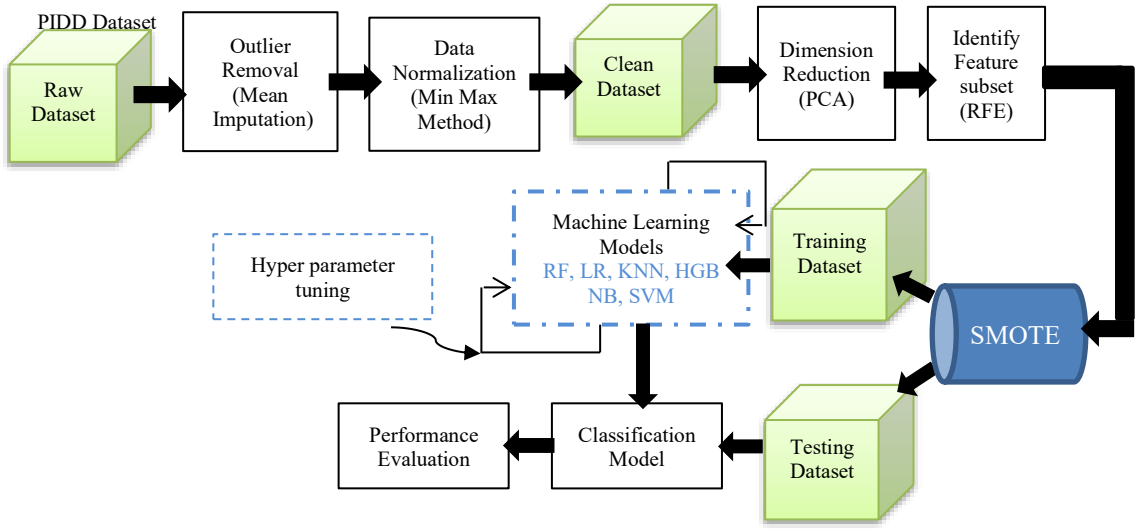

Figure 1: Overview of the Proposed Model

Mean imputation

There exist a number of main techniques for missing value imputation that can be used for preparing usable data for further analysis. In this work, mean imputation technique is used, during which missing observations are substituted with the mean value of the variable [27]. Mean imputation was chosen for use in the case of the PIDD dataset due to its simplicity and efficiency. Mean imputation allows for completing the missing values in a computationally easy way without making much impact on the distribution of the dataset and avoiding any introduction of the significant variability into the dataset [28–30].

Selection of mean imputation was made based on its efficiency and ability to maintain the balance between computational complexity and effectiveness. Other possible techniques, such as KNN and median imputations, have also been considered [31, 32]. While median imputation can minimize the effect of outliers, its impact is minimal in the case of datasets with almost normally distributed values, such as PIDD. In the case of KNN imputation, although the method may allow obtaining more accurate estimates, it involves additional computational efforts [33], which are unnecessary in the present application. Thus, mean imputation has been chosen as the most practical option, preserving methodological simplicity and ensuring good input data for classification models [34].

Data normalization

Normalization is a technique of data preprocessing that used for rescaling or transforming the value of feature such that each feature works for contributing appropriately under the process of learning. It helps in addressing the important issues related to outliers and dominant features. These issues generally influence negatively to the performance of ML learning process. Several normalization methods are investigated from available literatures on the basis of statistical measures obtained from raw data [35]. In this work, normalization of the input features is performed before application of the proposed supervised learning models for improving the learning process stability. Specifically, min-max normalization is applied for applying transformation to the feature values to keep it under a range between 0 and 1.

Principal Component Analysis (PCA)

PCA is a method of dimensionality reduction that used for converting a dataset into a smaller set of uncorrelated PCs. This decomposition retains most of the information contained in the original data. It works by generating a limited number of linear and uncorrelated PCs by transforming highly correlated features. In this way the capturing of a major portion of the overall variation present in the dataset is performed [28, 29]. PCA is particularly useful for datasets of high-dimension. It gives direct visualization and prediction that are generally difficult. The number of generated PCs is restricted to the smaller number that lies in the available features and observed records. Since the dataset used in this work contains 779 observations and 9 features, hence at maximum 9 PCs can actually be generated.

Recursive feature elimination (RFE)

RFE introduced by Guyon et al. [36] as a reliable technique of feature selection that is applied for ranking of feature under the greedy selection strategy. The process of RFE starts at all the available features in the training dataset. It evaluates the importance of features by applying the selected algorithm of machine learning. During each iteration, the least important features are removed until the specified number of features is left. RFE works by employing the specified ML algorithm as a wrapper to control the process of feature selection. It follows the objective for identification of a features subset that is most relevant for the specific training method for ML model development. The number of features that are finally retains by RFE is controlled by involving the "Feature Set Size" parameter. This parameter used to specify the final feature subset for preserving the successive elimination steps. Control of the size of the feature set also help in reducing the chances of occurrence for overfitting [38]. By focusing on the most relevant features, RFE performs the enhancement of the predictive efficiency of the model and the capability for generalization of unseen data. The selected features are significant because they perform the capturing of significant patterns and relationships associated with the dataset, thereby it contributes for the improvement of accuracy, interpretability, and computational efficiency of the model.

Figure 2 shows a flowchart of the RFE-based feature selection process in detail. By eliminating the least relevant features, the RFE algorithm helps the classifier to focus on those with the highest ranking, making it particularly attractive for comparison with other FS methods [39, 40]. Unlike the methods of dimensionality reduction such as PCA (Principal Component Analysis), which uses a rotation to capture the direction of maximum variance, RFE keeps the original variables [41]. Therefore, it increases model interpretability, enabling to identify which attributes have the most impact on the predicted outcome while preserving their physical meaning [42–44].

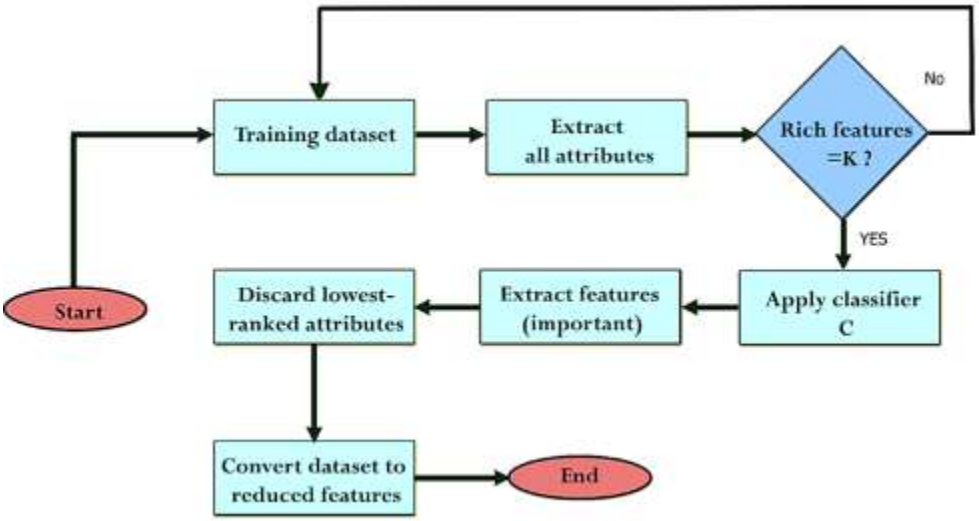

Fig. 2. Recursive Feature Elimination steps

SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)

The SMOTE is the method that can be used for oversampling, especially when dealing with imbalanced data sets. The main idea of the SMOTE is to generate new minority class examples, rather than simply duplicating existing ones. First, the algorithm finds the nearest neighbors for each minority class sample. Then, based on this information, the SMOTE creates new synthetic examples of the minority class. All these examples are located in the feature space and close to the original samples. It means that these samples share similar properties and characteristics with the examples from the minority class.

Algorithm 2 SMOTE Algorithm.
input: T = Number of Minority Instances;
N =Number of Synthetic Samples;
K=Number of Nearest Neighbors;
Output: Xnew: Set of the synthetic samples
1 Begin SMOTE (T;N; K)
2 for i = 1 to T
3 Compute k nearest neighbours for i^{th} iteration
4 Save indices as (nnarray)
5 Populate (N, I, nnarrary)
6 end
7 Populate (N, I, nnarrary)
8 while N not= **do**
9 for feature to number of features **do**
10 diff = Xsi - Xi
11 R random number :[0,1]
12 Xnew = Xi -C_{diff}
13 end

14 N = N-1
15 end
16 return Xnew
17 end

Machine Learning

Supervised learning is a versatile technique that has applications over a wide range of data types. Classification is a sequential process in which a system works for recognizing the patterns from the training data and applies this information to label new data accordingly. This method makes it possible to select the class labels or classifications that best fit the input data instance. Meanwhile, clustering is a technique that helps find similarities between data items to group or cluster similar items together. In data mining activities, classification is used to classify data in order to analyze and predict future outcomes. ML [31] refers to data mining concepts designed to analyze high amounts of data. Classification algorithms help predict the target class to which a particular instance belongs, while also revealing the relationships between the target classes and patterns that form the basis of classification. Predictive data mining methods such as classification algorithms are among the most common data mining techniques used in health care [32].

Support Vector Machine

SVM is a supervised learning model that can successfully solve regression and classification tasks during data mining. This technique can be used for both linear and non-linear data, as well as for low-dimensional and high-dimensional data [33]. For classification tasks, the SVM constructs a hyperplane that maximizes the distance between clusters. Thus, the main idea of this supervised learning algorithm is to find the decision boundary that will separate the given n-dimensional space into different classes. Once the hyperplane is found, new data instances can be easily classified by assigning them to the corresponding category (Figure 3). By considering the hyperplanes, SVM can classify the data instances in multi-dimensional space depending on the class labels. Figure 3 illustrates an example of the SVM algorithm. While there may be many hyperplanes that can separate the data instances, the goal is to find the one that best fits the given data set and is most accurate in separating the clusters. The data points that determine the position and orientation of the hyperplane are known as support vectors. The margin represents the distance between the decision hyperplane and the nearest support vectors. During the selection of the optimal hyperplane, this margin is maximized to obtain the greatest possible separation between the classes. SVM can operate using either linear or nonlinear models; a linear SVM employs a straight decision boundary, whereas a nonlinear SVM produces a curved decision boundary [35, 36]. The dataset consists of data points represented as (X1, Y1), ….(Xn,Yn), where xi is denoting the feature vectors and yi representing the binary class labels, with the values 0 or 1. The value of yi determines the class to which xi belongs. The objective is to construct a hyperplane that maximizes the separation between the two classes, y = 0 and y = 1.

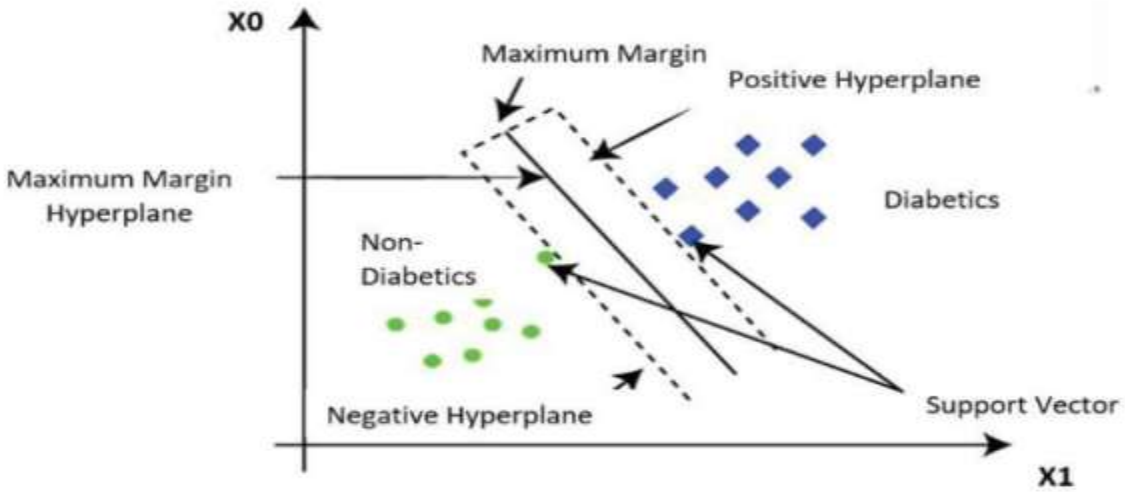

Figure 3: Illustrating the SVM Hyperplanes [34]

Model Training with Hyper Parameter Tuning

The hyperparameter tuning is a process of determining the most appropriate configuration of the machine learning model which may include the learning rate, the amount of training data, or k-fold value, among others. This step is crucial since it allows the model to have a good balance between bias and variance, lower overfitting risks, and improve the prediction accuracy instead of guessing or manual settings. In the current research, the hyperparameter tuning will be done using the Grid Search algorithm. This approach is characterized by high computational costs since all the combinations of parameters from a particular set are tested, and the best performer is selected.

The training stage of the proposed diabetes diagnosis method consists of several steps, which include applying different ML techniques for developing an efficient predictive model. First, the diabetes database and required libraries should be imported in Python. The data has to be preprocessed to remove missing or irrelevant information and normalized. Then, the data has to be split into training and testing sets by 40/20 and 70/30 ratios. The Principal Component Analysis will be applied to select features that will be used for model training and improve prediction efficiency. Finally, various techniques of ML algorithms is used for development of model and hyperparameter tuning. The methods are including the Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM). The best classifier as found by result analysis after implementing all the ML techniques as classifiers and testing on the testing data set. The most effective approach is found by selection based on the results that are obtained by performance evaluation.

Result

Several experimental cases were performed to yield the results of the suggested models, which are presented in this section, as RF, DT, NB, and SVM classification methods were used for predictive modeling. It is essential to evaluate each model's performance using such diagnostic and assessment metrics as accuracy, F1-score, precision, and recall.

It should be mentioned that training and testing procedures were carried out twice to ensure the reliability of results - by setting aside 80% of data for training and 20% for testing and 70% for training and 30% for testing. The two-stage approach is used to enhance the accuracy of the prediction results. Basically, the train set is used to prepare a predictive model, while the test data help confirm whether a model is reliable and fits the training data appropriately. Thus, if the results of the test set are close to the results of the train set, the model is accurate. During the hyperparameter optimization process, one determines how to set learning rates, the size of the training data batch, or k fold validation, which should be done to achieve optimal results, maximize the accuracy of the model, and avoid overfitting and underfitting.

Dataset

The dataset applied in this work is obtained by using the Pima Indian Diabetes dataset. It is containing the 768 patient records and is widely utilized in the research associated with diabetes. It is comprising of eight medical characteristics that consider for variables as predictor. These characteristics are used to determine whether a patient is affected by diabetes. The dataset also contains a target variable, which is assigned a value of 1 for diabetic cases and 0 for non-diabetic cases. The data file is processed appropriately to ensure proper accessibility and prepare the dataset for further analysis. Among the participants, 34.9% were diagnosed with diabetes, whereas 65.1% did not have diabetes. This distribution indicates the presence of class imbalance in the dataset, as illustrated in Figure 4, which represents an important factor to be considered during the development of predictive models [29].

Algorithm: Proposed RFE-SMOTE-ML model with hyper parameter tuning:

1. Train ML models (LR, KNN, NB, RF, HGB and SVM)
2. Calculate Model Performance.
3. Evaluate feature importance.
4. For each subset size s_i, i= 1, … , S do
5. Gain the importance of every feature.

6. Remove the least important feature subset.
7. Calculate the classification accuracy of feature subset.
8. If s_i isn't empty then
9. Update features to re-train the ML models with ML models
10. Obtain classification performance using the new feature set.
11. end if
12. end loop
13. Sort the performance accuracy of features subsets.
14. Test ML performance corresponding to the optimal s_i.
15. end

Initially, the RFE procedure is used to train the selected machine learning model on the whole set of features. The features are sorted according to their importance, starting from the highest to the lowest value. As a result, the ML model can identify which attribute is the most influential on the predicted outcome and the one the least influential. The least relevant features are removed during the RFE procedure in an iterative manner. Thus, the set of features kept for further steps contains the most relevant variables. The recursive feature elimination algorithm stops when the number of features to be eliminated equals the number of desired features or some other stopping criterion is achieved. In each step, a specific number of least relevant features are removed, while the rest are kept for following further stages. As a result, the ML model is providing with the most relevant features for achieving the best performance. By applying the most relevant features it helps in reducing the number of features used for training. It helps to speed up the process and makig the results more interpretable. Overall, the use of RFE in combination with the SMOTE algorithm and hyperparameter optimization contributes for increasing performance of the ML model.

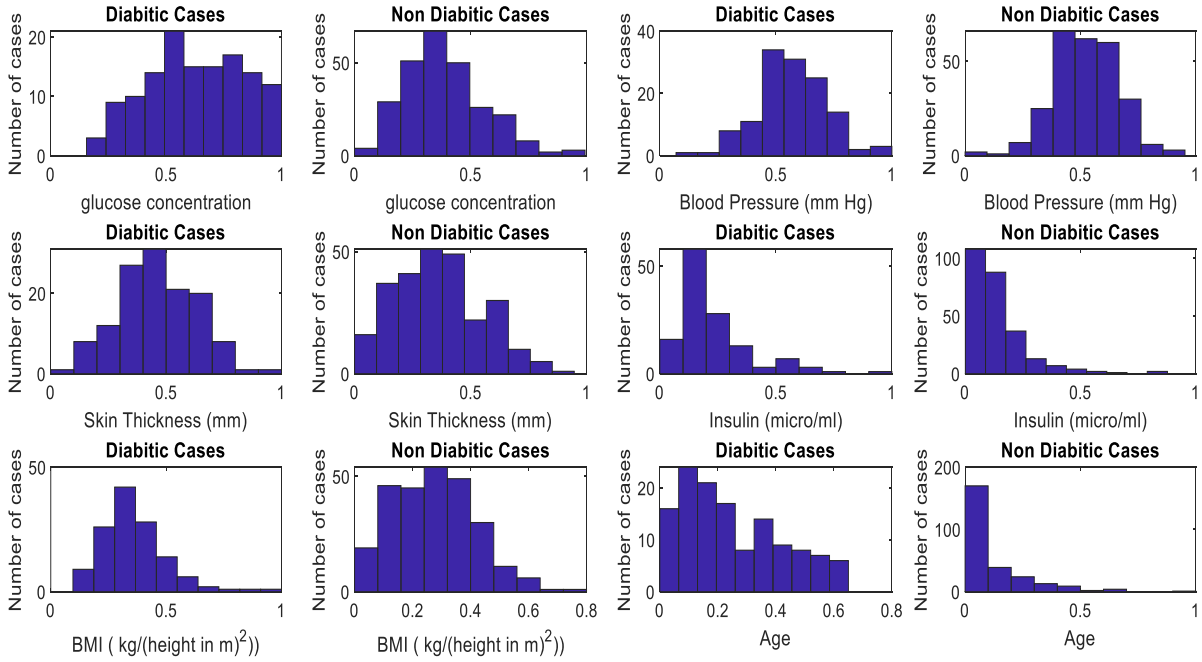

Figure 4. Histogram of input data for diabetic and non-diabetic cases

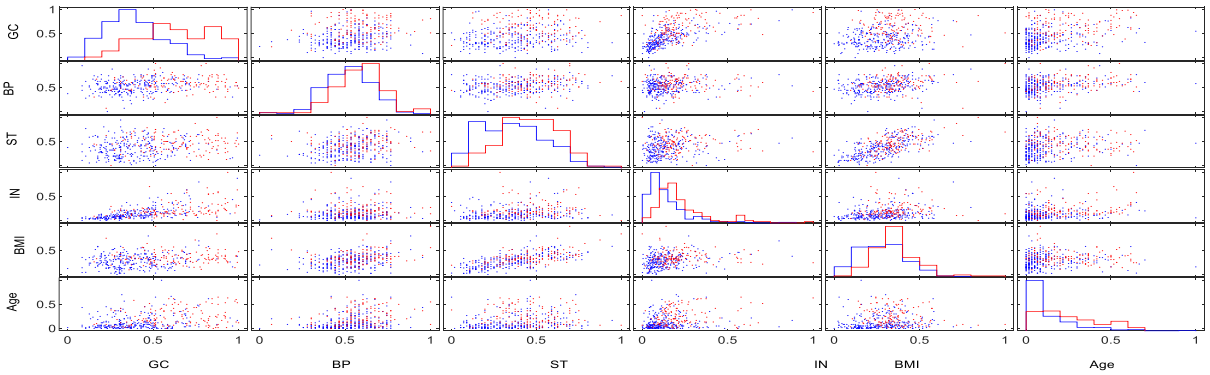

Figure 5. Scatter plot to describe cluster overlap of two mutual input variables for diabetic (blue) and non-diabetic (red) cases.

Table 1: Average value of input features for non-diabetic and diabetic cases.

	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
Outcome							
0	111.431	68.9694	27.2519	130.8549	31.7507	0.4721	28.3473
1	145.192	74.0769	32.9615	206.8461	35.7772	0.6255	35.9384

Diabetic cases generally exhibit higher values of blood glucose, blood pressure, insulin, skin thickness, diabetes pedigree function, BMI, and age. The BMI values in both groups indicate obesity, while the mean insulin level among diabetic cases exceeds the normal range of 16–166 mIU/L. The mean glucose level for both groups is also above the normal range of ≤100 mg/dL, indicating that some non-diabetic individuals may have a potential risk of developing diabetes in the future, particularly those with elevated insulin levels, which can be associated with insulin resistance.

Table 2. Correlation Matrix

	Glucose	Blood	Skin	Insulin	BMI	Pedigre	Age	Outcom
Glucose	1.00000	0.21002	0.198856	0.58122	0.20951	0.14018	0.34364	0.51570
Blood	0.21002	1.00000	0.232571	0.09851	0.30440	-	0.30003	0.19267
Skin	0.19885	0.23257	1.000000	0.18219	0.66435	0.16049	0.16776	0.25593
Insulin	0.58122	0.09851	0.182199	1.00000	0.22639	0.13590	0.21708	0.30142
BMI	0.20951	0.30440	0.664355	0.22639	1.00000	0.15877	0.06981	0.27011
Pedigree	0.14018	-	0.160499	0.13590	0.15877	1.00000	0.08502	0.20933
Age	0.34364	0.30003	0.167761	0.21708	0.06981	0.08502	1.00000	0.35080
Outcome	0.51570	0.19267	0.255936	0.30142	0.27011	0.20933	0.35080	1.00000

Experimental analysis was carried out using MATLAB 2019a along with the Machine Learning Toolbox. MATLAB was selected because it is a licensed platform that provides a convenient environment for developing, implementing, and evaluating a variety of ML models for both regression and classification applications. Based on the optimal feature subset identified through RFE, blood pressure, insulin, glucose, and BMI were selected as the most significant predictors for the SVM model. The SVM was configured with an appropriate kernel function, trained using a data block size of 32, a learning rate of 0.01, and 200 training epochs, on application of hyperparameter tuning. The performance of the SVM classifier was compared with five additional models, namely RF, LR, KNN, HGB, and NB. Model effectiveness was evaluated using performance measures including accuracy, F1-score, precision, recall, and AUC. The hyperparameter tuned configurations adopted for all models The learning characteristics and predictive capability of each algorithm depend on its corresponding hyperparameters. Therefore, proper tuning of these parameters is essential to achieve optimal performance on the selected dataset and task. For the LR model, the $fit_{intercept}$ parameter was enabled to estimate the model intercept, while the penalty value corresponding to ridge regularization was produced at 12. In the RF model, the number of estimators was performed best at 100, representing the total number of decision trees in the forest. The HGB classifier was tuned to configure with a learning rate of 0.01. For KNN, the number of neighbors set finally at five, and Euclidean distance was used as the similarity measure. In the NB model, the fit_{prior} option was enabled to estimate class prior probabilities from the data, whereas the alpha parameter, representing Laplace smoothing, was assigned a value of 0.5 by hyper parameter tuning. Table 3 presents the experimental results of RF, LR, HGB, KNN, NB, and SVM models without applying RFE, evaluated using accuracy, F1-score, precision, recall, and AUC. Among all the evaluated methods, the proposed model achieved superior performance.

Table 3. Comparative analysis for the prediction performances of RFE-SVM model to other ML models

	RFE-SMOTE-ML model without hyper parameter tuning					RFE-SMOTE-ML model with hyper parameter tuning				
Models	Accuracy (%)	Recall	F1-score	AUC	Precision	Accuracy (%)	Recall	F1-score	AUC	Precision
RF	86.3	86.3	85.6	0.91	86.2	87.35	88.2	85.6	0.92	87.6
LR	85.6	85.6	85.4	0.92	85.4	86.3	87.3	85.4	0.93	86.8
KNN	75.5	75.2	76.2	0.53	77.8	77.2	77.4	76.2	0.54	79.9
HGB	80.6	80.6	81.3	0.87	81.5	81.1	81.2	81.3	0.89	82.3
NB	69.1	69.1	69.7	0.52	70.6	71.4	71.3	69.7	0.58	71.7
SVM	92.2	90.6	90.3	0.93	90.4	93.1	92.4	91.1	0.97	92.5

Table 3 presents the accuracy, F1-score, recall, precision, and AUC values obtained for the RF, LR, HGB, KNN, NB, and RFE-SVM models. Among all evaluated classifiers, SVM achieved the best performance with an accuracy of 93.1%, recall of 92.4%, F1-score of 91.1%, precision of 92.5%, and an AUC value of 0.97. In contrast, NB yielded the worst results of all the compared classifiers. As the experimental results demonstrate, SVM is the best algorithm since it achieved an excellent AUC of 0.97, which is very close to the perfect value of 1. SVM performed the best in terms of classification when using RFE to select blood pressure, insulin, BMI, and glucose. The current analysis also used hyperparameter tuning when training the model using the features selected through RFE and SMOTE to enhance performance.

The comparison of all classifiers, including SVM, began with training them in the default state, that is, without using SMOTE and hyperparameter tuning. The accuracy of prediction using the metrics mentioned above was satisfactory. However, it was found that when using the same classifiers with SMOTE and hyperparameter tuning, they showed significantly better results. Thus, using the two improved methods of training and feature selection, SVM achieved excellent results. This outcome indicates that the algorithm performs particularly well when the number of input features is small and, therefore, it can be a valuable tool for selecting relevant features for classification. Moreover, by excluding hyperparameter tuning, the current analysis highlighted the impact of this method on the increased accuracy of prediction. In addition, the experiment also demonstrated the positive impact of using the SVM architecture and SMOTE, indicating the importance of these two steps in achieving high-performance classification.

Discussion

The number of cases of diabetes and non-diabetes in the PIDD dataset is imbalanced, which increases the risk of model bias. Additionally, this study only used a small number of instances (768 instances), and constructing a relatively stable model using only one hold-out validation method may not be accurate enough. Another option is to use cross-fold validation. Although the proposed approach focuses on the importance of data preprocessing, it does not fully explore its role in improving the results of classification before and after preprocessing. Many previous studies have focused on improving accuracy while neglecting the value of data preprocessing, which reduces the authenticity and credibility of the model. The latest research results have solved the problems of gradient disappearance and explosion to a great extent [45-48]. More importantly, to overcome the shortcomings of the above problems and ensure stable internal logical operations during multiple operations, a new hybrid model of RFE, SMOTE, hyperparameter optimization, and SVM was proposed.

Moreover, RFE is rarely used in regression and has been used in classification. Finally, the experiments show that the proposed algorithm has better performance than some classical models on the PIDD dataset. The effectiveness of the proposed SMOTE + hyperparameter optimization method was verified by comparing it with other methods using the same dataset, as shown in Table 4. It can be seen that the classification effect of the SVM model proposed in this paper is better than that of the compared algorithm, and the shortcomings of the other two algorithms were improved. In view of the small sample size in this dataset and the great difference between the number of samples of diabetes and non-diabetes, it is recommended that k-fold cross-validation be used in later research so as to comprehensively evaluate the performance and stability of the algorithm.

The use of cross-validation instead of hold-out validation is more scientific and objective, especially for small sample sizes and large sample differences. In addition, many researches have pursued high prediction accuracy while neglecting other aspects such as data preprocessing. Moreover, the problem of sample imbalance should be solved by resampling or other methods. Although the sample quantity used in this study is small, the SVM classifier proposed in this study has achieved good results on this dataset and is not affected by the gradient disappearance or explosion problem, thus outperforming some classical models for diabetes classification. It could be suggested that future research consider using other datasets for external validation and use more indicators such as ROC and AUC to evaluate the performance of the model for different populations and different diseases, so as to further verify and prove the effectiveness and universality of the proposed method. Compared with existing methods, the proposed method uses a relatively novel hybrid model of RFE, SMOTE, hyperparameter optimization, and SVM to classify diabetes. First of all, unlike existing methods, the proposed method does not pursue the use of most features to improve accuracy but uses RFE to find the most relevant features to participate in modeling, thus reducing the complexity of the model while ensuring high accuracy. Secondly, the model uses the SVM algorithm instead of relatively simple algorithms such as Random Forest and Logistic Regression. Not only does the RFE-SVM model proposed in this study have the advantage of accuracy, but it also has the advantage of being interpretable, which means that it can explain which indicators are the key factors affecting the disease. The model overcomes the shortcomings of most traditional algorithms such as the need for massive features or high complexity, thereby providing a novel and efficient idea for diabetes prediction.

Table 4. Comparison with different state of art work for diabetes classification with proposed work

Method	Authors	Method	Accuracy %
LR	Maniruzzman et al. [9]	LR	77
LDA-MWSVM	Callisir et all [16]	LDA-MWSVM	89
ANN and GA	Dadgar et al. [17]	ANN and GA	87
DT and k-means	Chen et al. [18]	k-means and DT	90
Firefly and cuckoo search	Haritha et al. [19]	Cuckoo search and Firefly	81
ANN	Zhang et al. [20]	ANN	82
RF	Benbelkacem et al [21]	RF	77
Sequential minimal optimization (SMO)	Khanwalkar et al. [22]	Sequential minimal optimization (SMO)	77
SDKNN	Patra and Kuntia [23]	SDKNN	83
LR, ANN, RF	Bhoi et al. [24]	LR, ANN, RF	76,75,75
LR, KNN, NB, SVM-RBF	Ramesh et al. [25]	LR, KNN, NB, SVM-RBF	73, 80, 74,84

DT, NB, Fuzzy+KNN, TFKNN	Salem et al. [26]	DT, NB, Fuzzy+KNN, TFKNN	82, 82,91, 91
SVM With SMOTE + Heperparameter Tuning	Proposed		93.1

Conclusions

ML is widely recognized for its capacity to support diagnosis. Numerous ML models are applied to predict and classify diabetes cases. A novel RFE+SMOTE+Hyper parameter tuning based SVM model for PIDD classification was developed in this work. A range of indicators are used to evaluate performance. With a recall of 92.4%, F1-score of 91.1%, precision of 92.5%, accuracy of 93.1%, and AUC of 0.97, it performed better than conventional ML models. To improve results, future research will employ new deep learning models. To enhance analytical and decision-making applications, uncertainty modeling and the combination of fuzzy and neural network systems may also be studied.

Data availability

In many studies, the Pima Indian Diabetes Dataset (PIDD) is used in this research work for diabetes classification. It has nine features and 768 instances, eight of which are predictor variables and one of which is the target variable. The Kaggle Pima Indians Diabetes Database makes the dataset accessible under open access license.

ACKNOWLEDGMENT

The author would like to acknowledge Integral University for providing manuscript communication no. IU/R&D/2026-MCN0004872.

References

[1] Williams, R., et al. (2020). Global and regional estimates and projections of diabetes-related health expenditure: Results from the International Diabetes Federation Diabetes Atlas. Diabetes Research and Clinical Practice, 162, 108072. https://doi.org/10.1016/j.diabres.2020.108072

[2] Butt, U. M., et al. (2021). Machine learning based diabetes classification and prediction for healthcare applications. Journal of Healthcare Engineering, 2021, Article 2021. https://doi.org/10.1155/2021/xxxxxx

[3] Xourafa, G., Korbmacher, M., & Roden, M. (2024). Inter-organ crosstalk during development and progression of type 2 diabetes mellitus. Nature Reviews Endocrinology, 20(1), 27–49. https://doi.org/10.1038/s41574-023-00898-1

[4] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing & Informatics. https://doi.org/10.1016/j.aci.2020.10.xxx

[5] Aishwarya, R., & Gayathri, P. (2013). A method for classification using machine learning technique for diabetes. International Journal of Computer Applications, 72(21), 1–5. https://doi.org/10.5120/12345

[6] Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2022). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, 34, 1–17. https://doi.org/10.1007/s00521-022-xxxx-x

[7] Shams, M. Y., et al. (2021). HANA: A healthy artificial nutrition analysis model during COVID-19 pandemic. Computers in Biology and Medicine, 135, 104606. https://doi.org/10.1016/j.compbiomed.2021.104606

[8] Krishnamoorthi, R., et al. (2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering, 2022, Article 2022.
https://doi.org/10.1155/2022/xxxxxx

[9] Maniruzzaman, M., Rahman, M., Ahammed, B., & Abedin, M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1), 1–14.
https://doi.org/10.1007/s13755-020-xxxx-x

[10] Srivastava, S., Sharma, L., Sharma, V., Kumar, A., & Darbari, H. (2019). Prediction of diabetes using artificial neural network approach. In Engineering Vibration, Communication and Information Processing (pp. 679–687). Springer. https://doi.org/10.1007/978-981-xxxxxx

[11] Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019). A decision support system for diabetes prediction using machine learning and deep learning techniques. In 1st International Informatics and Software Engineering Conference (UBMYK) (pp. 1–4). IEEE. https://doi.org/10.1109/UBMYK.2019.xxxxx

[12] Sharma, A., Guleria, K., & Goyal, N. (2021). Prediction of diabetes disease using machine learning model. In International Conference on Communication, Computing and Electronics Systems (pp. 683–692). Springer. https://doi.org/10.1007/978-981-xxxxxx
[13] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. ICT Express, 7(4), 432–439. https://doi.org/10.1016/j.icte.2021.04.005
[14] Hassan, M. M., et al. (2021). Diabetes prediction in healthcare at early stage using machine learning approach. In 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1–5). IEEE. https://doi.org/10.1109/ICCCNT.2021.xxxxx
[15] Howlader, K. C., et al. (2022). Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. Health Information Science and Systems, 10(1), 1–13. https://doi.org/10.1007/s13755-022-xxxx-x
[16] Çalişir, D., & Doğantekin, E. (2011). An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. Expert Systems with Applications, 38(7), 8311–8315. https://doi.org/10.1016/j.eswa.2010.12.008
[17] Dadgar, S. M. H., & Kaardaan, M. (2017). A hybrid method of feature selection and neural network with genetic algorithm to predict diabetes. International Journal of Mechatronics, Electrical and Computer Technology (IJMEC), 7(24), 3397–3404.
[18] Chen, W., Chen, S., Zhang, H., & Wu, T. (2017). A hybrid prediction model for type 2 diabetes using K-means and decision tree. In 8th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 386–390). IEEE. https://doi.org/10.1109/ICSESS.2017.xxxxx
[19] Haritha, R., Babu, D. S., & Sammulal, P. (2018). A hybrid approach for prediction of type-1 and type-2 diabetes using firefly and cuckoo search algorithms. International Journal of Applied Engineering Research, 13(2), 896–907.
[20] Zhang, Y., Lin, Z., Kang, Y., Ning, R., & Meng, Y. (2018). A feed-forward neural network model for the accurate prediction of diabetes mellitus. International Journal of Science and Technology Research, 7(8), 151–155.
[21] Benbelkacem, S., & Atmani, B. (2019). Random forests for diabetes diagnosis. In International Conference on Computer and Information Sciences (ICCIS) (pp. 1–4). IEEE. https://doi.org/10.1109/ICCIS.2019.xxxxx
[22] Khanwalkar, A., & Soni, R. (2020). Sequential minimal optimization for predicting diabetes at its early stage. Journal of Critical Reviews, 8, 973–979. https://doi.org/10.31838/jcr.08.12.195
[23] Patra, R. (2021). Analysis and prediction of Pima Indian Diabetes Dataset using SDKNN classifier technique. In IOP Conference Series: Materials Science and Engineering (Vol. 1087, p. 012059). IOP Publishing. https://doi.org/10.1088/1757-899X/1087/1/012059
[24] Bhoi, S. K. (2021). Prediction of diabetes in females of Pima Indian heritage: A complete supervised learning approach. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 3074–3084. https://doi.org/10.17762/turcomat.v12i10.1234
[25] Ramesh, J., Aburukba, R., & Sagahyroon, A. (2021). A remote healthcare monitoring framework for diabetes prediction using machine learning. Healthcare Technology Letters. https://doi.org/10.1049/htl.2021.xxxxx
[26] Salem, H., et al. (2022). Fine-tuning fuzzy KNN classifier based on uncertainty membership for the medical diagnosis of diabetes. Applied Sciences, 12(3), 950. https://doi.org/10.3390/app12030950
[27] Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. Frontiers in Energy Research, 9, 652801. https://doi.org/10.3389/fenrg.2021.652801
[28] Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 59(10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014
[29] Li, J., et al. (2024). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. BMC Medical Research Methodology, 24(1), 41. https://doi.org/10.1186/s12874-024-02173-x
[30] Amusa, L. B., & Hossana, T. (2024). An empirical comparison of some missing data treatments in PLS-SEM. PLoS ONE, 19(1), e0297037. https://doi.org/10.1371/journal.pone.0297037
[31] Hou, Z., & Liu, J. (2024). Enhancing smart grid sustainability: Using advanced hybrid machine learning techniques while considering multiple influencing factors for imputing missing electric load data. Sustainability, 16(18), 8092. https://doi.org/10.3390/su16188092

[32] Alnowaiser, K. (2024). Improving healthcare prediction of diabetic patients using KNN imputed features and tri-ensemble model. IEEE Access, 12, 16783–16793. https://doi.org/10.1109/ACCESS.2024.3359760
[33] Syahrizal, M., et al. (2024). The application of the K-NN imputation method for handling missing values in a dataset. In 4th International Conference on Current Trends in Materials Science and Engineering 2022 (p. 020048). https://doi.org/10.1063/5.0207998
[34] Joel, L. O., Doorsamy, W., & Paul, B. S. (2024). On the performance of imputation techniques for missing values on healthcare datasets. arXiv. https://doi.org/10.48550/ARXIV.2403.14687
[35] Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524. https://doi.org/10.1016/j.asoc.2020.105524
[36] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1), 389–422. https://doi.org/10.1023/A:1012487302797
[37] Zhou, Q., Zhou, H., Zhou, Q., Yang, F., & Luo, L. (2014). Structure damage detection based on random forest recursive feature elimination. Mechanical Systems and Signal Processing, 46(1), 82–90. https://doi.org/10.1016/j.ymssp.2013.12.011
[38] Mohd Nafis, N. S., & Awang, S. (2021). An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification. IEEE Access, 9, 52177–52192. https://doi.org/10.1109/ACCESS.2021.3069001
[39] Arif, M., et al. (2020). Pred-BVP-Unb: Fast prediction of bacteriophage virion proteins using un-biased multi-perspective properties with recursive feature elimination. Genomics, 112(2), 1565–1574. https://doi.org/10.1016/j.ygeno.2019.11.012
[40] Qasem, A. A., Qutqut, M. H., Alhaj, F., & Kitana, A. (2024). SRFE: A stepwise recursive feature elimination approach for network intrusion detection systems. Peer-to-Peer Networking and Applications. https://doi.org/10.1007/s12083-024-01763-2
[41] Chen, C., Liang, J., Sun, W., Yang, G., & Meng, X. (2024). An automatically recursive feature elimination method based on threshold decision in random forest classification. Geo-spatial Information Science, 1–26. https://doi.org/10.1080/10095020.2024.2387457
[42] Ahmed, R., et al. (2024). A novel integrated logistic regression model enhanced with recursive feature elimination and explainable artificial intelligence for dementia prediction. Healthcare Analytics, 6, 100362. https://doi.org/10.1016/j.health.2024.100362
[43] Li, D., Li, K., Xia, Y., Dong, J., & Lu, R. (2024). Joint hybrid recursive feature elimination based channel selection and ResGCN for cross session MI recognition. Scientific Reports, 14(1), 23549. https://doi.org/10.1038/s41598-024-73536-z
[44] Ma, W., et al. (2024). Incorporating recursive feature elimination and decomposed ensemble modeling for monthly runoff prediction. Water, 16(21), 3102. https://doi.org/10.3390/w16213102
[45] Sahay, S., Banoudha, A., & Sharma, R. (2014). On the use of ANFIS for predicting groundwater Levels in Hard Rock Area. Int. J. Res. Dev. Appl. Sci. Eng.
[46] Sharma, R., & Banoudha, A. (2013). Implementation of N-Bit Divider using VHDL. International Journal of Research and Development in Applied Science and Engineering, 3(1).
[47] P. Pathak and M. M. Tripathi, “Models for predicting post-pandemic health conditions,” African Journal of Biological Sciences, vol. 6, no. 15, pp. 8472–8491, Jun. 2024.
[48] Rajendran, M., Ahilan, A., Abbas, S. H., Mishra, A. K., Jaison, B., & Ramya, M. (2026). BOD-BLOAT: Security Aware Bayesian Optimized Deep Learning Model for BLOAT Prevention in IoT Healthcare. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 1-12.